Editorial

The American Statistical Association statement on *P*-values explained

For better or worse, Null Hypothesis Significance Testing (NHST) with its associated "*P*" values has become the standard for most published medical literature. However, "*P*" values are difficult to understand and interpret, even for established researchers. This has led to a lot of unfavorable attention to this issue in the recent past,[1] especially in the context of research misconduct. It has become the perennial butt of scientific cartoonists such as Randall Munroe at http://xkcd.com [Figure 1].[2] It is in this background that for the first time in its 177 year history, the American Statistical Association released a "Statement on Statistical Significance and *P*-values" with six principles. Wasserstein and Lazar have explained the context, process, and purpose of this statement in The American Statistician.[3]

As practicing physician-scientists, it is important for us to understand the context and "significance" of this statement. This editorial attempts to explain the salient features of this statement from the perspective of Indian anesthesiology research, based on the explanations provided by Wasserstein and Lazar.

Let me postulate a specific clinical research scenario for this purpose. A new antiemetic Nopov has been tested against placebo in a sample of patients undergoing day care gynecological surgery. The number of patients vomiting on the first postoperative day was lower in the treatment group (45/100) compared to the placebo group (60/100) with a *P*-value of 0.03.

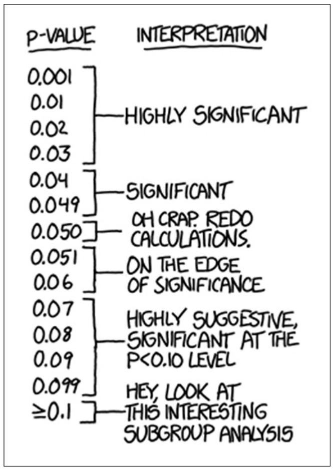
P-VALUE
INTERPRETATION
0.001
0.01
0.02
0.03
HIGHLY SIGNIFICANT
0.04
0.049
SIGNIFICANT
0.050
OH CRAP. REDO CALCULATIONS.
0.051
0.06
ON THE EDGE OF SIGNIFICANCE
0.07
0.08
0.09
0.099
HIGHLY SUGGESTIVE, SIGNIFICANT AT THE P<0.10 LEVEL
≥0.1
HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS

Figure 1: "If all else fails, use 'significant at $P > 0.05$ level' and hope no one notices." (http://xkcd.com/1478/, Randall Munroe, Creative Commons Attribution-NonCommercial 2.5 License)

Principle 1: P-values can indicate how incompatible the data are with a specified statistical model.

A *P*-value is one of the ways of summarizing the incompatibility between the observed data and a proposed model for the data. The most common model we use, is the so-called "null hypothesis," which in practice essentially proposes that Nopov has no effect whatsoever. The smaller the *P*-value, the larger the incompatibility of the data with the null hypothesis. A *P*-value of 0.03 says that 45% patients in the Nopov group will vomit by chance in 3% of samples drawn from a population, in which Nopov has no effect. It does not say anything about the population in which Nopov has an effect.

Principle 2: P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

A *P* value of 0.03 does not mean that the probability of Nopov not having any effect is 3%. Neither does it mean that 45% of patients in the Nopov group vomited by chance. As the last sentences in Principle 1 explain, a *P*-value is a statement about the data in the context of a null hypothesis. It does not say anything about the null hypothesis or the alternate hypothesis. It is not an error probability.

Principle 3: Scientific conclusions and business or policy decisions should not be based only on whether a P-value passes a specific threshold.

The results of this study should not lead you to conclude that Nopov reduces postoperative nausea. A conclusion is not binary yes/no because of one study. Researchers should consider the context of the study while deriving scientific inferences. These should include the design of the study (e.g., improper randomization or poor concealment of allocation, leading to selection, or other types of biases), the quality of measurements (e.g., if vomiting is not documented in real time, but is based on recollection at an interview conducted a week later, leading to recollection bias), the external evidence for the phenomenon under study (e.g., studies showing that Nopov does not penetrate the blood–brain barrier, or that it does, or that it causes profound sedation), and the validity of statistical assumptions that underlie the data analysis. A low *P*-value should never be the sole basis for a scientific claim.

Principle 4: Proper inference requires full reporting and transparency.

All the analyses done should be reported fully. Conducting multiple statistical tests on the same data and reporting only those with low *P*-values makes the reported analysis uninterpretable. For example, in the given study, the number of patients vomiting may have been analyzed for the period that the patient was in the postanesthesia care unit, or the first 6, 24, 48 or 72 h. The number of episodes of vomiting per patient in all these given periods could have been tested instead of the number of patients. That gives us ten possible tests. If all ten are conducted, and only the ones with a $P < 0.05$ are reported, as commonly happens, it amounts to research misconduct going by various names such as cherry-picking, data dredging, p-hacking, significance chasing, or more politely "selective inference." This is one of the main causes of the spurious excess of statistical significance in published literature. "Valid scientific conclusions based on *P*-values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted, and how those analyses (including *P*-values) were selected for reporting."[3]

Principle 5: A P-value, or statistical significance, does not measure the size of an effect or the importance of a result.

The threshold of statistical significance that is commonly used is a *P*-value of 0.05. This is conventional and arbitrary. It does not convey any meaningful evidence of the size of the effect. A *P*-value of 0.01 does not mean the effect size is larger than with a *P*-value of 0.03. The *P*-value would have been 0.000002, if we had sampled 1000 patients instead of 200 in the Nopov study and obtained the same results (i.e., the same effect size). Similarly, if an effect is measured to a high enough precision, the *P*-value will change. For example, if in the above study, the incidence of postoperative vomiting was 46%, the *P*-value would have been 0.047, which is considered statistically significant. If we then measure the incidence more precisely and get a value of 46.4%, the *P*-value would become 0.054, considered nonsignificant. A similar change can take place in the opposite direction as well.

Statistical significance does not automatically equate to scientific, human, or economic significance. It might be that even if Nopov does, in fact, reduce the incidence of vomiting by 15%, this might not be clinically relevant. It might not matter to the patient, if, for example, it produces severe dysphoria which is more uncomfortable for the patient. It might not make economic sense to use the drug, if, for example, it costs 10,000 rupees to treat one patient.

Principle 6: By itself, a P-value does not provide a good measure of evidence regarding a model or hypothesis.

A context-less *P*-value without any other evidence provides very limited information. A large *P*-value is not evidence of your alternate hypothesis since an arbitrarily large number of other hypotheses are consistent with the observed data. Data analysis should not conclude with the calculation of the *P*-value. Correct and careful interpretation of statistical tests requires examining the sizes of effect estimates and confidence limits as well as precise *P*-values. Other approaches may provide a more direct evidence of the size of an effect or the correctness of a hypothesis albeit with further statistical assumptions. These approaches include methods emphasizing estimation over testing, Bayesian methods, likelihood ratios, and others.

The fundamental problem discussed here is the implicit practice of defining success on passing an arbitrarily defined threshold. If this is followed, biases will occur regardless of whether the threshold being considered is a *P*-value, a 95% confidence interval, Bayes factor, false discovery rate, or any other measure. It is better to promote transparency in study design, conduct, and reporting than to rely on a single binary criterion of whatever type.

At the present time, as responsible scientists, we should do the following at a minimum. In the above-mentioned study, we should specify what exactly the null hypothesis was and what the alternate hypothesis. We should specifically document the effect size that we considered clinically and economically important and relevant to the patients in question. We should calculate the sample size required based on this effect size, and appropriate α and β values, which may not be 0.05 and 0.20 as in most studies. We should define in adequate detail measures to reduce bias such as choosing an appropriate population, randomization, and concealment of allocation. Moreover, we should implement them during the conduct of the study. We should document beforehand the outcomes of interest, and the methods and time of their measurement. If we want to perform NHST, we should define *a priori* what significance tests we want to perform on what data. We should interpret the results in context, with an understanding of the underlying phenomena, with complete reporting of all the analyses performed. Finally, we should supplement the data summaries and the *P*-values with estimates of the effect sizes with measures of their uncertainty, and other methods such as likelihood ratios, confidence, credible, or prediction intervals.

Lakshmi Narayana Yaddanapudi

Department of Anaesthesiology and Intensive Care,
PGIMER, Chandigarh, India

Address for correspondence: Prof. Lakshmi Narayana Yaddanapudi,
Department of Anaesthesiology and Intensive Care, PGIMER,
Chandigarh - 160 012, India.
E-mail: narayana.yaddanapudi@gmail.com

References

1. Siegfried T. Odds are, it's wrong: Science fails to face the shortcomings of statistics. Sci News 2010;177:26. Available from: https://www.sciencenews.org/article/odds-are-its-wrong. [Last accessed on 2016 Nov 1].
2. Licensed Under a Creative Commons Attribution-Non Commercial 2.5 License. Available from: http://www.xkcd.com. 1478/. [Last accessed on 2016 Nov 1].
3. Wasserstein RL, Lazar NA. The ASA's statement on P-values: Context, process, and purpose. Am Stat 2016;70:129-33. Available from: http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN. [Last accessed on 2016 Nov 11].

Access this article online

Quick Response Code:

Website:
www.joacp.org

DOI:
10.4103/0970-9185.194772

How to cite this article: Yaddanapudi LN. The American Statistical Association statement on *P*- values explained. J Anaesthesiol Clin Pharmacol 2016;32:421-3.